Assignment-based Subjective Questions

---

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: There were more bike rentals in the year 2019 compared to 2018. Bike rentals are likely to be high in summer and fall seasons. Bike rentals are highest when the weather is clear or partly cloudy and lowest when it is snowy or rainy.

---

2. Why is it important to use drop_first=True during dummy variable creation?

Ans: For representing categorical variables with n-levels, n-1 columns are sufficient as dummy variables. Pandas function get_dummies() is used to create dummy variables from categorical variables. By default, it will give n dummy variables for n-levels in a categorical variable. Since we need only n-1 variables to represent n-levels, get_dummies() with drop_first=True should be used. Having one less variable also helps in building a model with less correlation between variables.

---

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: The temperature (temp) variable has the highest correlation with the target variable.

---

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: To validate linear relationship between dependent and independent variables and little or no multicollinearity being present, I checked p-score and VIF. To check if error terms are normally distributed, I plotted a histogram of error terms.

---

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Temperature, year and winter are the top 3 features that explain the demand of the shared bikes.

General Subjective Questions

---

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is an algorithm that provides linear relationship between a dependent and independent variables to predict values of dependent variable. It is a statistical method used in data science and machine learning for predictive analysis.

An independent variable is a variable that remains unchanged or independent due to the change in other variables. But, the dependent variable changes with fluctuations in independent variables. The linear regression model attempts to predict the value of the dependent variable.

So, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables.

---

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

---

3. What is Pearson's R?

Ans: Pearson's R (also known as Pearson's Correlation Coefficient) is used to establish a linear relationship between two variables. It gives an indication of the measure of strength between two variables and the value of the correlation coefficient can be between -1 and 1.

---

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Usually dataset contains features that vary in magnitudes, units and range. If scaling is not performed, the algorithm only takes magnitude in account and not units hence resulting in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

The difference between normalized scaling and standardized scaling is that normalization brings all the data points in a range between 0 and 1 and standardization replaces the values by their Z scores, bringing all the data points into a standard normal distribution which has mean as zero and standard deviation as one.

---

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: The value of VIF is infinite when there is a perfect correlation between two independent variables. The R-squared value is 1 in this case and as per VIF's formula = 1/(1 – R2), the value of VIF becomes infinite. This concept suggests that there is multicollinearity between variables and one of them needs to be dropped for defining a working regression model.

---

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: The quantile-quantile (Q-Q) plot is used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.
